Capstone Project Proposal Template

Notes:

- This should take no more than one hour to complete – the clearer you are about the business problem you're working to solve with your ML-driven solution, the easier your proposal will be to complete
- This will be uploaded to your repo, which will be a part of your final submission
- Due date for submission is 1/16

Instructions:

1. Download this document as a Word Doc
2. Answer each question using a few sentences, at most
3. Save your completed proposal as a PDF
4. Create a project GitHub repo (if you have yet to do so)
5. Add your instructor as a collaborator (username dodgy719) to your project repo
6. Add your mentor as a collaborator
7. Push your proposal PDF (created in Step 3) up to your repo
8. Copy the URL corresponding to the location of the PDF in your repo
9. Submit the copied URL using this link

Housing Blue Book

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?
    - How to use patient data to predict whether or not an individual has or is at risk for heart disease.
- What industry/realm/domain does this apply to?
    - This AI model applies to the medical industry
- What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)
    - The motivation behind this project is to provide the general population with a tool that helps indicate whether someone has or is at risk for heart disease. Heart disease is the leading cause of death in the United States.

Data Understanding

- What data will you collect?
    - I will be collecting a dataset that contains patients' vital information relevant to indicators of heart disease. (https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?select=heart.csv)
- Is there a plan for how to get the data (API request, direct download, etc.)?

    - Direct download
- What are the features you'll be using in your model?
  - Age
  - Sex
  - chest pain type (4 values)
  - resting blood pressure
  - serum cholestoral in mg/dl
  - fasting blood sugar > 120 mg/dl
  - resting electrocardiographic results (values 0,1,2)
  - maximum heart rate achieved
  - exercise induced angina
  - oldpeak = ST depression induced by exercise relative to rest
  - the slope of the peak exercise ST segment
  - number of major vessels (0-3) colored by flourosopy
  - thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?
    - Deciding which features correlate best that serve as indicators for someone who is at risk for or has heart disease.
- What are some of the cleaning/pre-processing challenges for this data?
    - Removing irrelevant features/variables, some variables are scaled differently (normalization), and some columns need to be renamed or dropped.

Modeling

- What modeling techniques are most appropriate for your problem?
    - Linear regression- how closely are certain features related to whether or not someone has heart disease.
- What is your target variable? (remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target)
    - Whether or not someone has heart disease (0 or 1)
- Is this a regression or classification problem?
    - This is a classification problem

Evaluation

- What metrics will you use to determine success (MAE, RMSE, Accuracy, Precision etc.)?
    - Accuracy- given a list of vitals information for any patient- how accurate can we predict that any individual has or is at risk for heart disease.

Tools/Methodologies

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?

- Random Forest- Identify trends of heart disease and risk factors for heart disease.